DOCUMENT RESUME

ED 174 686 **TM 009 658**

AUTHOR	Nassif, Paula M.; And Others
TITLE	Generating Parallel Test Forms for Minimum Competency Exams.
PUB DATE	Apr 79
NOTE	19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Complexity Level; *Criterion Referenced Tests; *Equated Scores; Functional Illiteracy; Higher Education; *Item Analysis; Mathematical Models; Minimum Competency Testing; Senior High Schools; State Programs; Teacher Certification; *Test Construction; Testing Programs; Test Items; *Test Reliability; Tests
IDENTIFIERS	Florida; Georgia; *Parallel Forms Reliability; Test Security

ABSTRACT

A procedure which employs a method of item substitution based on item difficulty is recommended for developing parallel criterion referenced test forms. This procedure is currently being used in the Florida functional literacy testing program and the Georgia teacher certification testing program. Reasons for developing parallel test forms involve test security and avoidance of the spread of knowledge about test content resulting from repeated administrations. Both maintenance of strict content parallelism and individual examinee classification consistency are insured when parallel forms are generated using p-values for item-by-item substitution. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174686

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Generating Parallel Test Forms for Minimum Competency Exams

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Paula M. Nassif

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paula M. Nassif
Paul D. Pinsky
Sherry A. Rubinstein

National Evaluation Systems, Inc.
Amherst, Massachusetts

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1979.

TM009 658

Generating Parallel Test Forms for
Minimum Competency Exams

A great deal of planning and anticipation of impact and change has been devoted to the design and initiation of minimum competency testing programs. Issues—political, legal and technical in nature—have been confronted to varying degrees of success. By following some programs from inception through initial administration and on through successive cycles, one can amass a tremendous amount of information about minimum competency programs. The well-known minimum competency program in Florida is one that has been carefully watched.

Because most minimum competency programs are ongoing, the cyclical repetition of activities has both provided the researcher with the opportunity to be more and more efficient, and in many cases, more insightful. It is however in these implementation phases that several testing issues arise. The educational use of minimum competency testing identifies specific psychometric concerns that must be met for continued successful program implementation. One issue of critical importance, and the focus of this research, is that of generating parallel forms of criterion-referenced tests used for minimum competency testing.

In continuing programs, information about the content of the test and in some instances the actual questions on the exam are recalled following an administration and passed along to future examinees. In addition, in some programs candidates are allowed or required to retake

the exam until a passing score is achieved. Due to repeated administrations, a single candidate could receive a passing score on the test more as a function of his/her knowledge of the test than knowledge of the content domain being assessed.

There are other situations in which parallel forms of a minimum competency test must be available. If a test were stolen and all or sections of the test were distributed, then a parallel form would have to be prepared and ready for use prior to the next scheduled administration. For different reasons, the copy of a minimum competency exam may have to be made available. Some state policies indicate that for the assistance in development of local minimum competency programs, the test administered statewide be available following administration. Finally, in the case of a legal challenge, a test may be required to be available for inspection. In all of these cases, test security is threatened.

In summary then, there are two primary reasons for developing parallel test forms: one is to maximize test security, and the other is to minimize advance knowledge of the content of the test due to repeated administrations.

At least two current testing programs address this issue. They are the functional literacy program in the state of Florida and the Georgia Teacher Certification Testing Program.

The overriding concern in parallel forms development in both cases is consistency of classification from administration to administration, using tests that are content-referenced to the same set of objectives (Rentz, 1978). In minimum competency programs, insurance of consistent

correct classification across different forms of the test is required, so that passing or failing is not a function of "which form you took" or "when" you participated. Further, these exams are not isolated components but in most cases are part of a program which includes preparation, instruction, testing and remediation. The tests are content domain-referenced (if not objective-referenced) and therefore parallel tests must conform not only psychometrically but content-wise as well.

There are two approaches to solving the problem of ensuring classification consistency and content parallelism in the development of minimum competency tests. Simply stated, one method is to equate scores, and the other is to equate test forms.

An example of the first method is the approach used by Educational Testing Service for calibrating College Board Scores (Angoff, 1968); and two examples of the second are the Rasch model and item substitution by p-values (percent correct or item difficulty). The goal in all three cases is the same: consistency of score or score classification across test forms for an examinee of given ability. The definition of parallel test forms used to measure minimum competency is that all examinees have the same probability of passing both test forms. This paper discusses the parameters of item substitution in Florida and Georgia, presents the methods used, and the results of parallel form development by p-value.

Parameters

Some salient parameters which influence score equating and development of parallel forms in both programs named above are: the nature of criterion-referenced tests, the distribution of scores in minimum competency exams, the pass/fail cutoff, and the amount of item overlap across administrations. These parameters were carefully considered in the process of selecting the method of developing parallel forms. The discussion which follows briefly describes the main features of the approaches under consideration.

In the Angoff method, raw scores are converted to scale scores so that the emphasis is on correct score conversion. Scores are calibrated to adjust for the achievement level of the candidate population by using a set of items common to both forms of the test. By regressing one set of scores on another (a simple example is a linear transformation), scores can be calibrated using the information obtained from the common equating items (Angoff, 1968).

According to Wright (1968), Rasch models calibrate test items independent of the examinee sample used for calibration purposes. Further, the measurement of examinees occurs independent of the test items used for measurement purposes. Because items are calibrated 'sample-free", equating of parallel test forms occurs directly from the construction and calibration of the items.

Both Florida and Georgia utilize criterion-referenced tests, which typically do not coincide with an important classical test theory assumption (Glaser, 1963; Hambleton, et al., 1978); that is, they typically generate

scores that are skewed to the high end of the scale. It is unclear from current research whether Angoff and Rasch methods of equating work with highly skewed distributions. Furthermore, in Georgia, adequate numbers of examinees are not available to produce reliable item calibrations. The factor of insufficient numbers of examinees (on which reliable item calibrations can be made) greatly inhibits the possibility of using the Rasch model here (Whitely, 1977).

In both the Georgia and Florida programs, examinees receive only a pass or fail classification. However, the Georgia classification is based simply upon total test score, while the Florida classification is based upon passing several objectives within the test as well as passing the total test. This simplifies typical score equating because the consistency of decision is focused on ensuring that the examinee would receive the same *classification* (pass or fail) on administrations with parallel forms. In other equating situations (e.g., SATs), the candidate's score must be scaled so that he/she has some (acceptable) probability of obtaining the *same score* on administrations with parallel forms.

Item Substitution Method

Florida Functional Literacy Test

The statistical problems involved in developing parallel forms of the Florida Functional Literacy Test (FFLT) are distinctive. The use of the FFLT is not to measure a student's actual ability level relative to functional literacy, but rather to determine if a student has achieved a minimal level of functional literacy as determined by the "passing" criteria. This distinction is discussed by Cronbach and Gleser (1965).

In order to define fully the statistical problems of developing a parallel form for the FFLT, notation must first be defined.

Let

S = total number of test items on the test

s = total number of correct answers on the test by a student

A = test Form A (the old or existing test)

B = test Form B (the new parallel test which may contain some of the same test items as test Form A)

α = the achievement or knowledge of a student

Then

$P_A\{s|\alpha\}$ = probability that a student with achievement α will score s on test A for $s = 0,\ldots,S$

Note that

$$\sum_{s=0}^{S} P_A\{s|\alpha\} = 1$$

Let M be the result that a student has mastered or passed the test. There are two criteria for mastering or passing the test. These are:

1. the student gets at least C items correct on the test (Mathematics or Communications)

2. the student shows mastery on at least D objectives defined within the test

Let M_T be the result that the student gets at least C items correct on the entire test, and let M_{OB} be the result that a student has mastered an objective on the test for OB=1,...,NOB, where NOB is the number of objectives on the test.

A sufficient condition to define that test Form B is parallel to test Form A is:

1. $P_B\{M_T|\alpha\} = P_A\{M_T|\alpha\}$ where $P_A\{M|\alpha\} = \sum_{s=C}^{S} P_A\{s|\alpha\}$

and

2. $P_B\{M_{OB}|\alpha\} = P_A\{M_{OB}|\alpha\}$ for OB=1,...,NOB

Note that these conditions are not necessary to define parallel tests since a student need only master D objectives of the NOB objectives on the test. One could define necessary conditions in terms of the combinatorics of mastering D of the NOB objectives.

At this point the problem is reduced to solving conditions 1 and 2. Assume that both Test Form A and test Form B contain S items. These

items are referenced by

$$I_A^1 \ldots . I_A^S \text{ for test Form A}$$

and

$$I_B^1 \ldots . I_B^S \text{ for test Form B}$$

Note that the above rationale is symbolic, so that I_A^8 may be the identical test item to I_B^6.

Let

P_A^S = difficulty level of an item I_A^S, s=1,...,S

P_B^S = difficulty level of an item I_B^S, s=1,...,S

These difficulty levels are defined relative to the population that responded to test Form A at the first administration. Thus, we know the values of P_A^S s=1,...,S, while the values of P_B^S s=1,...,S are unknown. Items that potentially could be used on test Form B were field tested to estimate the parameters P_B^S.

The field test yielded estimates of item difficulty for new potential test items. Thus, any item put on Form B will contain an estimate of its difficulty P_B^S. (Note that many of the items on Form B will be the same as items on Form A, thus P_B^S will be an exact estimate.)

Now, a sufficient condition for satisfying conditions 1 and 2 is

3. $P_B^S = P_A^S$, s=1,...,S

Thus, the rule for generating Form B so that it is parallel to Form A is to substitute only those items that have the same estimated difficulty level.

A practical consideration is that the field test may not generate items of the exact same difficulty level. Indeed there is a sampling error associated with the estimation of the difficulty level of the new test items. Therefore, one should balance the difficulty level of new test items across the *set* of new test items within a given objective. For example, if one new item for an objective is slightly more difficult than the old ones, it should be balanced by another new item that is slightly less difficult.

Estimation of Difficulty of New Test Items

The new test items that could be used on the Florida parallel test Form B were field tested to gather estimates of P_B^S. The original test form was taken by eleventh-graders in November 1977. The new test items were field tested on twelfth-graders in March 1978. Part of the statistical problem is to estimate the new test items' difficulty relative to the original eleventh-grade sample. This was accomplished by a technique known as "seeding" or linking some of the old test items to each set of new items being field tested. Each objective was field tested with 10 new items and 4 old items per objective. The objectives were split across several test forms so that each set of 5 new items was seeded with at least 2 old items.

There are several methods that may be used to translate the field test results to estimates of difficulty of the new test items on the eleventh-grade population. The simplest is to assume a linear shift in difficulty for all items or for items within an objective. Another approach

is to use the Rasch model (Wright, 1977). However, as presented later, the Rasch approach was not appropriate for the characteristics of the Florida situation.

The sampling plan used to gather the field test data was a stratified cluster sampling plan with 800 students per item. The stratification was based upon size of district and ethnic origin, while the clustering involved testing intact classrooms. The sampling plan used generated estimates of item difficulty ±5% at the 95% level of confidence.

In the analyses of data, the Rasch model did not work satisfactorily for the item substitution process. The data below present some of the problems encountered.

TABLE I

RASCH ESTIMATES OF ITEM DIFFICULTY

Item	Form X		Form Y		Form Z	
	P	D	P	D	P	D
A	94	-.40	95	.12	96	-.24
B	95	-.77	96	.01	92	.31
C	92	.00	94	.31	96	-.51
D	91	.08	93	.54	96	-.48
E	92	-.41	90	.74	85	1.12
F	89	.11	91	.75	86	1.06
G	90	.01	88	.95	90	.85
H	88	.31	95	.27	91	.59
I	92	-.38	96	-.62	96	-.71
J	95	-1.02	98	-.87	98	-1.00

Form X is the original test administered to eleventh-graders in November 1977, Form Y and Form Z are two field test forms administered in March 1978.

Items A-J are 10 items that appeared on all three test forms. P is the item difficulty stated as the percentage of students who correctly answered the item while D is the Rasch estimate of item difficulty. The Rasch estimates have not been corrected for the scaling constant. An examination of the data will show that this was not possible.

The Rasch analysis which is supposed to be "person-free", should have enabled the researchers to estimate the difficulty of the new items (field tested in March 1978) using twelfth-graders on the eleventh-graders tested in November 1977. The reasons the Rasch model was not well suited to this application are:

a) the Rasch model is not applicable for very easy or very hard test items. A usual procedure in Rasch analysis is to eliminate all test items with difficulty levels above .90 or .95. Clearly this rule cannot be used for minimum competency testing programs.

b) the Rasch model assumes a unidimensional scale. Thus items A and C becoming easier over time, while item E becoming more difficult does not fit the Rasch model. The analysis of the data showed a definite shift in difficulty among objectives over time. Moreover, any proficiency testing program that bases the passing score on several objectives clearly does not meet the unidimensional assumptions of the Rasch model.

Georgia Teacher Certification Testing Program

The Georgia Teacher Certification Testing Program consists of criterion-referenced tests in 17 teaching fields administered four times a year. Initial teaching certification in Georgia is based on three criteria: (1) grades in selected college courses, (2) teaching skills observed during student teaching, and (3) minimum competence in the teaching field content knowledge. An important aspect of the testing

program is the development of parallel test forms that can be administered over time. The expected number of examinees taking each exam ranges from over 1,000 in one field to under 50 in another field. The problem is to develop test forms that are statistically parallel, yet different enough to maximize test security and minimize advance knowledge of the content of the test.

The design was to construct 150 item-tests in each field. These 150 items were selected by objective according to a job analysis study (Nassif, 1978). The 150 items in each field are divided into 5 subtests as shown below.

Subtests

Scorable equating	= 37 items
Scorable replaceable A	= 25 items
Scorable replaceable B	= 25 items
Scorable replaceable C	= 25 items
Experimental	= 38 items

Each of the subtests is a stratified sample of all items on the test. The pass/fail decision is based upon the 112 scorable test items. Thus the 38 experimental items do not contribute to the examinee's score. When a new form is developed, all scorable replaceable A items are deleted from the test and 25 of the 38 experimental items are used as the new scorable replaceable A items. The remaining experimental items are also deleted from the test. The substitution of experimental items into scorable replaceable A items is done within an objective for items

of the same difficulty level as determined from the previous testing session. This item substitution plan preserves the content validity of the test as well as the statistical difficulty of the test. In addition, 38 new experimental items are written for objectives measured by items in the scorable replaceable B subtest for the next round of item substitution.

An obvious problem with the above plan is that one cannot always identify test items of the exact same difficulty level for substitution. Therefore, one must attempt to average the differences in item difficulty over subareas within the teaching field and also finally across all 25 items being substituted. This latter process can easily be accomplished.

Verifying Parallelism

The critical question that must be raised and answered is: How does one know that the two forms of the test are parallel? Further, what statistical test can one use to measure the parallelism of the test forms. The true definition of parallel test forms is that the probability distribution of score results is the same for the same examinee on both test forms. Knowing that the distribution of an examinee's test scores approximates a normal distribution, one can test the equality of test forms by simulating an examinee's responses, generating two sample distributions, and testing the equality of variances (assuming the means are equal) using an F-test.

Consider the following hypothetical example of five test forms of 150 items each. Items 1-75 have the same p-value. Items 76-150 have the same p-value, which is different from that of Items 1-75. Each test has the same expected score of 75 correct answers. However, the variance of the expected distribution varies greatly in each case.

TABLE II

Simulation of Scores for Five Test Forms

FORM	P_1	P_2	MEAN	VARIANCE	STANDARD DEVIATION
A	.01	.99	75.1	1.2	1.1
B	.10	.90	74.8	12.9	3.6
C	.25	.75	74.9	24.9	5.0
D	.40	.60	74.7	32.2	5.7
E	.50	.50	74.9	45.2	6.8

P_1 = p-value of items 1-75.

P_2 = p-value of items 76-150.

The mean and variance of the 5 forms were generated by simulating 400 examinee responses for each form and computing the sample mean and variance. An F-test clearly shows that all five forms are not parallel. One can see the lack of parallelism if one considers a passing score of 80 out

of 150 items on the test. Virtually 0% would pass Form A, while 17% would pass Form C, and approximately 26% would pass Form D.

Next consider some actual data from the Georgia Teacher Certification Testing Program. Forms A and B each contain 150 items; Form B contains 25 experimental items from Form A as scorable in place of 25 scorable replaceable items on Form A.

TABLE III

Comparison of Test Forms

FORM	MEAN	VARIANCE	F-RATIO
1A	83.4	18.4	1.01
1B	83.4	18.6	
2A	78.3	18.9	1.03
2B	78.3	18.4	
3A	80.5	18.8	1.05
3B	80.6	19.8	

The sample mean and sample variance were calculated by simulating 1,000 examinee responses to each form. Clearly the F-ratio supports the equality of distributions for the candidate and thus the parallelism of test forms.

These examples show that the procedure recommended for item substitution does in fact produce parallel test forms. That is, one can see that the score distributions for a candidate are the same across forms of the test, with item substitutions having been made on the basis of p-values.

Summary and Recommendations

Two primary considerations in generating parallel tests for Florida and Georgia were the maintenance of strict content parallelism and individual examinee classification consistency. Both of these were preserved by generating parallel forms using p-values as a basis for item-by-item substitution.

This paper

- presented the theoretical framework for considering a p-value substitution method
- described results from alternative approaches
- and presented data showing that items can be substituted within the content of an objective for items of the same difficulty, and balanced first within the objective and finally within content subareas to generate a parallel test form.

These results show that item substitution by p-values is a successful and efficient method of generating parallel tests for minimum competency programs. This procedure does maintain consistency of examinee score classification (pass/fail). As stated earlier, the inclusion of "equating" or "linking" items on all forms of the test is recommended for observing achievement shifts.

Both the Georgia and Florida projects involve criterion-referenced minimum competency tests. While both tests are similar in purpose (teacher licensing/high school diploma award), many other parameters differ (age of examinees, content of test, participation). The responsiveness of this approach to these programs attests to its potential for expanded further use.

References

Angoff, W.H. How we calibrate College Board scores. College Board Review, Summer 1968, 68, 11-14.

Benson, J., Crocker, L.M., and Ware, W.B. A comparison of three types of item analyses in test development using classical and latent trait methods. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada, March 1978.

Cronbach, C.J. and Gleser, G.C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Hambleton, R.J., Swaminathan, H., Algina, J. and Coulson, D. Criterion-referenced testing and measurement: a review of technical issues and developments. Review of Educational Research, Winter 1978, 48, No. 1, 1-47.

Nassif, P.M. Standard setting for criterion-referenced licensing tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada, March 1978.

Rentz, R.R. Monitoring the quality of an item pool calibrated by the Rasch model. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada, March 1978.

Whitely, S.E. Models, meaning and misunderstandings: some issues in applying Rasch's theory. Journal of Educational Measurement, 1977, 14, 227-235.

Wright, B.H. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J., Educational Testing Service, 1968.

Wright, B.H. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.